Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Multi-agent Reinforcement Learning for Travel Time Reliability

Ioannis Christofilogiannis

Supervisor: Prof. Vinny Cahill

August 13, 2026

A dissertation submitted in partial fulfilment of the requirements for the degree of M.Sc. Computer Science (Intelligent Systems)

Abstract

Between its ramps, a motorway has no signals, meters, or signs: nothing for learning-based traffic control to actuate. Two decades of reinforcement learning for traffic have therefore lived in the infrastructure, while the road between the ramps stays uncontrolled. This dissertation moves the learning to the vehicles. It models road space as a conveyor of moving slot reservations and treats every vehicle as a recurrent learning agent choosing among five discrete slot manoeuvres: hold the current reservation, shift forward or backward, or change into a compatible slot in an adjacent lane. On top of a rule-based slot simulator and the SUMO microscopic traffic simulator, a complete multi-agent learning system is built: lifetime-scoped agent rows, compact local observations with optional task signals, an 18-term reward with hybrid per-agent credit assignment, and a three-layer safety envelope that makes collisions structurally impossible. Multi-Agent Proximal Policy Optimization (MAPPO) and Independent PPO (IPPO) are trained on a segment suite inspired by Dublin's M50, spanning straight and curved corridors, passable hazards, lane drops and gains, on-ramp merges, and exit diverges, each evaluated against matched do-nothing and random-valid baselines.

Learning pays exactly where coordination is genuinely needed. On free-flowing roads the slot layer alone sustains throughput, and training adds discipline: vehicles stay within one slot of their assignment 96% of the time, using a quarter of the maneuvers of a random policy, and the behavior transfers zero-shot to an unseen curved corridor. Where scenarios demand more, trained vehicles take their assigned exit 88% of the time against an 18% baseline. In the on-ramp scenario they finish merges in their assigned mainline lane 72% against 42% while cutting the merge queue by up to 90%. They also clear a lane drop into the surviving lane 88% against 65% with a travel-time tail less than half as long. Every result, for every policy, at every demand level, is collision-free.

The most important finding is that IPPO consistently beat centralized MAPPO, which looks like evidence against centralized training with decentralized execution. Diagnosis traced the gap to the critic's input: a 1289-value joint observation, two thirds padding, in which the vehicle being valued is marked by a single flipped bit. Replacing it with Minimal MAPPO, a 29-value agent-specific critic holding only the global traffic state and the focal vehicle's own view, closes the gap on the merge, overtakes IPPO on the exit corridor. It also degrades most gracefully under demand far beyond training, with 26% less queueing than IPPO at the heaviest demand tested. Minimal MAPPO manages to match or beat IPPO on difficult scenarios because the critic can clearly see whose return it is predicting, due to the more targeted observation vector, while it is utilizing more observation features than IPPO. Finally, a native macOS alternative for SUMO GUI was built when the official interface failed on the test system. This work is a demonstration that, inside a safe environment, learning vehicle-level cooperation on a motorway is feasible.